

RESPONSE SURFACE METHODS FOR FORCE BALANCE CALIBRATION MODELING

Peter A. Parker
Research Scientist
NASA Langley Research Center
Hampton, Virginia 23681

Richard DeLoach
Senior Research Scientist
NASA Langley Research Center

Abstract

A "modern design of experiments" (MDOE) approach to balance calibration at NASA Langley Research Center focuses on the application of formal experimental design techniques to address weaknesses in the current calibration methodologies. The Single-Vector Balance Calibration System (SVS) has been developed as an innovative mechanical load application system specifically designed for the efficient and accurate execution of a formal experimental design. This paper emphasizes practical applications of response surface methodology with the analyses of experimental data. Calibration experimental design concepts including the estimation of the required data volume and an evaluation of the model prediction capability are presented. Randomization, replication, and blocking are proposed as means of tactical defense against systematic errors present in all calibration systems. Response surface methods are implemented in obtaining an adequate model with the minimum number of terms and partitioning of the unexplained variance. A systematic approach to augmenting a second order model with higher order terms is discussed. Applying formal experimental design techniques to force balance calibration provides a suite of sophisticated and elegant tools that advance balance calibration technology.

Background

Direct force and moment measurement of aerodynamic loads is fundamental to wind tunnel testing at NASA Langley Research Center (LaRC). Force balances are the state-of-the-art instrument that provides these measurements in six degrees of freedom. Electrically measured strain, as a function of load, forms the basic concept of force balance measurements that has been generally used since the 1940s. Ideally, each balance signal would respond only to its respective component of load, and it would have no response to other components of load. This is not entirely possible even though balance designs are optimized to minimize these

undesirable interaction effects. Ultimately, a calibration experiment is performed to obtain the necessary data to generate a mathematical model.

Over the past 60 years, there have been improvements in many areas of force balance technology, but relatively little has changed in the area of force balance calibration methodology. Calibration is the most critical phase in the production of a high quality force transducer. The goal of a calibration experiment is to derive a mathematical model that is used to estimate aerodynamic loading incurred during wind tunnel testing. The accuracy of this model is also determined during the calibration experiment. The present experimental approach is based on a one-factor-at-a-time (OFAT) methodology, where each independent variable is incremented individually throughout its full-scale range, while all other variables are held at a constant magnitude.

Calibration models are based on a polynomial equation where the balance response is a function of the applied load. This model can be thought of as a Taylor's series approximation to a general function. For example, with $k = 2$ design variables, a general polynomial can be expressed by:

$$\begin{aligned}
 f(x, \beta) &= \beta_0 && \text{(y - intercept)} \\
 &+ \beta_1 x_1 + \beta_2 x_2 && \text{(linear terms)} \\
 &+ \beta_{12} x_1 x_2 && \text{(interaction terms)} \\
 &+ \beta_{11} x_1^2 + \beta_{22} x_2^2 && \text{(pure quadratic terms)} \\
 &+ \beta_{111} x_1^3 + \beta_{222} x_2^3 + \beta_{112} x_1^2 x_2 + \beta_{122} x_1 x_2^2 && \text{(cubic terms)} \\
 &+ \beta_{1111} x_1^4 + \dots && \text{(quartic terms)} \\
 &+ \text{etc.},
 \end{aligned}$$

where k is the number of independent variables, x_i is the i^{th} independent variable, and β represents the coefficients in the mathematical model. Typically, the higher the degree of the approximating polynomial, the

more closely the Taylor series expansion will approximate the true mathematical function.¹ In the current LaRC calibration model, a degree of two is used, and therefore a second-order model is generated. For balance calibration there are six design variables, and the second order model for each response contains a total of 28 terms.

In order to set the independent variables of applied load, a high-precision mechanical system is required. Manual dead-weight balance calibration stands have been in use at LaRC since the 1940's. These simple methodologies produce high confidence results, but the process is mechanically complex and labor-intensive, requiring three to four weeks to complete.²

Over the past decade, automated balance calibration systems have been developed. In general, these systems were designed to automate the manual calibration process. Unfortunately, the automation of this tedious manual process results in an even more complex mechanical system that is quite expensive. Also, compared to manual systems, the quality of load application is deteriorated.³

There are a number of weaknesses in the current calibration methodology and available load application systems. In regard to methodology, the OFAT approach has been widely accepted because of its inherent simplicity and intuitive appeal to the balance engineer. LaRC has been conducting research in a "modern design of experiments" (MDOE) approach to force balance calibration. Formal experimental design techniques provide an integrated view to the calibration process. This scientific approach applies to all three major aspects of an experiment; the design of the experiment, the execution of the experiment, and the statistical analyses of the data.

Load application systems, both manual and automated, also have weaknesses. The manual systems, although generally considered accurate, are slow and tedious and provide many opportunities for systematic error. Automated systems that greatly reduce calibration time include additional sources of systematic error due to their mechanical complexity, and their expense makes them prohibitive for wide spread use. Both of these mechanical systems were designed around the OFAT calibration requirement to set independent variables one at a time and to obtain maximum efficiency of data collection.

In order to apply formal experimental techniques, a new mechanical system was required. An innovative approach to balance calibration has been developed at LaRC that integrates a unique load application system with formal experimental design methodology. The

Single-Vector Balance Calibration System (SVS) enables the complete calibration of a six component force balance with a single force vector.⁴ A primary advantage to this load application system is that it improves on the "trusted" aspects of current manual calibration systems. The SVS enables the efficient execution of a formal experimental design, is relatively inexpensive to manufacture, requires minimal time to operate, and provides a high level of accuracy in the setting of the independent variables. A photograph of the SVS is provided in Figure 1.

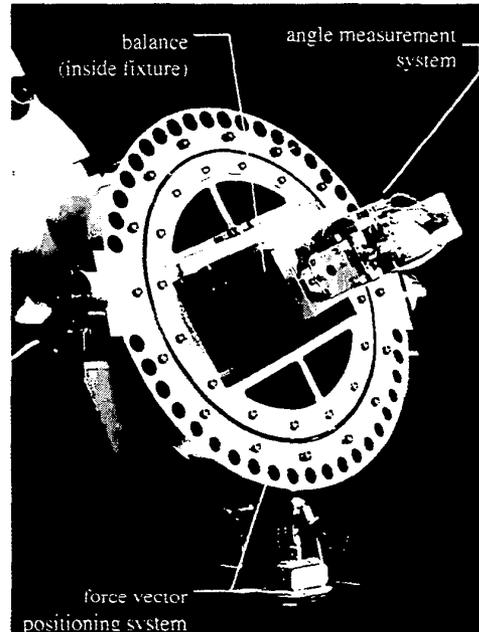


Figure 1. Single-Vector System

The SVS allows for single vector calibration, meaning that single, calibrated dead-weight loads are applied in the gravitational direction generating six component combinations of load relative to the coordinate system of the balance. By utilizing this single force vector, load application inaccuracies caused by the conventional requirement to generate multiple force vectors are fundamentally reduced. The angular manipulation of the balance, combined with the load point positioning system, allows the uni-directional load to be used to produce three force vectors (normal force, axial force, side force) and three moment vectors (pitching moment, rolling moment, yawing moment), with respect to the balance moment center. As a result, the use of a single calibration load reduces the set-up time for the randomized multi-axis load combinations required to execute a formal experimental design. Further details on the Single-Vector System are available in Reference 4. This paper is focused on the

analytical techniques that this new mechanical calibration system has enabled.

Introduction

This paper presents the application of formal experimental design principles to the balance calibration experiment. First, the factors involved in determining an adequate calibration design and its evaluation are presented. The calibration design is commonly referred to as the load schedule, but the term design will be subsequently used because it implies a process with clear objectives and criteria, not simply a list of load combinations. Balance engineers have posed questions about how to evaluate a calibration design. For example, how many data points are required in the calibration design? How can we evaluate the quality of a design before we get experimental data? Is there any way to measure or defend against systematic errors that are present in all calibration systems? These questions are addressed in the first section of this paper on the experimental design. Objective techniques are presented that enable the balance engineer to develop better, more efficient, experimental designs.

In the second section of this paper, the application of response surface methods (RSM) to balance calibration modeling is presented. Questions have been posed about how the analysis should be performed, and the interpretation of the results. For example, isn't the best model the one with the most terms? How do we know when a model coefficient is due to random noise instead of a real effect? How do we determine whether the model is adequate, and when higher-order terms are required? These questions are addressed in the RSM section of this paper. A formal experimental design enables the application of powerful analysis techniques that provide new insights into calibration results. These insights empower the balance design engineer to objectively analyze and report the results.

Experimental Design

Data Volume Estimation

The OFAT calibration design, that specifies the independent variables to be applied, was created to enable the determination of the mathematical coefficients using graphical techniques. Historically this was required due to the computational resources available in the 1950's, when this design was developed.⁵ Since this time, the design has remained nearly unchanged. In the LaRC OFAT design, there are a total of 81 load sequences performed sequentially in time. Each load sequence consists of a tare point, four increments, three decrements, and return tare point

providing a total of nine data points per sequence; 729 points in all.

Are 729 data points enough, not enough, or too many? How do we determine how many data points are required in a calibration design? The minimum number of points in an experimental design is bounded by the number of parameters in the model. For a d^{th} order polynomial model in k variables the number of parameters, p , can be determined according to the following equation,

$$p = \frac{(d + k)!}{d!k!} \quad (1)$$

For a second order model, there are a total of 28 parameters, or terms, in the model. This includes the intercept, six linear, six pure quadratic, and 15 two-way interactions. This means that there must be at least 28 distinct combinations of the independent variables in the calibration design. A design with 28 points is referred to as saturated, because there would be no additional degrees of freedom to assess the quality of fit of the model. In other words, all 28 points would lie on the calculated response surface. It is generally accepted that this is unsatisfactory in determining the model, but how many additional points are required? An objective approach to determine the total number of points required, referred to as the data volume, is presented in the next section.

The volume of data required in an experimental design depends on four primary parameters: 1) the repeatability of the measurement environment, 2) the precision requirement, 3) the inference error risk, and 4) the number of parameters in the model, as discussed above. The repeatability of the measurement environment is a function of how repeatable the independent variables can be set, and how the balance responses can be measured. It is a measure of the variance experienced in the calibration experiment and is determined by performing genuine replicates during the experiment. A genuine replicate is different from simply holding the independent variables at a constant setting and recording multiple data points. It requires changing the independent variables between identical set-point conditions. Prior to the execution of the design, this variance is based on historical data for a particular calibration system.

The precision requirement is commonly thought of as the required balance accuracy, or the desired quality of the prediction from the math model. When the calibration is completed, a calibration equation that predicts the balance response for a given set of applied loads is provided, this is the mathematical model. The

prediction of the model will not be the exact input loads unless the calibration is perfect. The precision requirement indicates how close is "close enough." For example, if a true normal force of 100 pounds is applied, how much different from 100 pounds is an acceptable prediction from the calibration model? In other words, when a 100 pounds is applied to the balance, we want the balance to read "100 pounds \pm X". What is "X"? This precision requirement should be tied directly to the wind tunnel test objectives.

There are two components of the inference error risk.⁶ First, there is the risk that the calibration model will predict a load that differs from the actual applied load by more than the specified precision requirement. The maximum acceptable probability of committing this type of error is generally represented by α . There is also a risk of failing to detect a true incremental change in load that is large enough to be important. The smallest such change in load defines the resolution requirement for the experiment. The maximum acceptable probability of committing this type of error, failing to detect an important change in load, is commonly denoted by β .

Typically a 95% confidence interval (0.05% risk) is reported from a balance calibration, which is an α type inference error risk for each component. This says that we will accept a 5% chance (1 in 20) that a freshly calibrated balance will predict a response that differs from the measured response in each component by more than the amount specified as the precision requirement. For example, if a 100 pound normal force is applied, and we have specified that we require the balance model to predict a response of 100 ± 0.08 pounds 95% of the time, so that only 5% of the time the balance will give a value either greater than 100.08 or less than 99.92 pounds.

A more appropriate specification is that there would be no more than a 5% chance of any of the six balance components be in error by more than the amount specified by the precision requirement. The probability of any one component being within the specification is computed such that the probability to the 6th power is no less than 0.95. Therefore, that probability follows.

$$P = (0.95)^{\left(\frac{1}{\delta}\right)} = 0.9915 \quad (2)$$

If there is a 99.15% probability that any one component is within specification, there is a 95% probability that all six components are within specification. This is what is intended when we specify a 95% confidence in the calibration result. (If we accepted 95% probability of being within specification on any one component,

the probability of at least one of the six being out of specification would be better than one chance in four).

Combining these parameters with the form of the mathematical model and the constraints involved in the execution of the design, provide the necessary information to define the required data volume. The purpose of determining the data volume is to "scale the experiment" to meet the objectives. In other words to quantify the volume of data required to meet the precision requirements specified, with the level of confidence required, in the presence of the amount of response variance that is anticipated.

One equation that can be used to estimate the required volume of data, N , is,⁷

$$N = p(z_\alpha + z_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2 \quad (3)$$

where p is the number of parameters in the model, z_α and z_β are the z-statistics associated with the inference error risks, σ is the standard error in the response measurements, and δ is the precision requirement.

As an example of this analysis, a typical LaRC balance will be used, designated as UT-39A. Its specifications are provided in Table I. The data volume for the normal, axial, and pitch components are computed; similar analyses can be performed for the other components.

Component	Design Load (pounds or inch-pounds)	Output (microvolts per volt)
Normal Force	150	1,645
Axial Force	30	1,956
Pitching Moment	200	1,433
Rolling Moment	30	1,494
Yawing Moment	100	1,399
Side Force	75	1,421

Table I. Balance UT-39A Specifications.

In order to determine the required data volume, the values of the repeatability of the measurement environment (expressed as the standard deviation), the precision requirement, and the inference error risks were estimated. The repeatability of the measurement environment was estimated at 0.4 microvolts per volt, based on historical calibrations. This number was converted into engineering units of the UT-39A and is supplied the first row of Table II.

	Normal	Axial	Pitch
σ	0.036	0.006	0.056
δ	0.150	0.030	0.200
α error risk	0.0085	0.0085	0.0085
β error risk	0.0170	0.0170	0.0170
z for both risks	2.632	2.632	2.632

(units for σ and δ are in pounds or inch-pounds)

Table II. Design values used to estimate data volume.

For this example the precision requirement was selected to be 0.10% of full-scale. This percentage was converted into engineering units for both components and is supplied in the second row of Table II. The acceptable inference error risk for α and β was specified at 0.0085 and 0.0170 respectively.

The z value for both inference error risks is 2.632. (They are the same because z_α is associated with a "two-sided" distribution, z_β is associated with a "one-sided" distribution, and the inference risks have a 2:1 ratio.) This can be found in standard statistical reference tables. In this case, the z -statistic was used because the repeatability estimate came from historical experience involving a statistically significant volume of data.

Substituting the above values into Equation 3 provides a relationship between the number of parameters in the model and the number of data points required. This relationship is simplified in Equation 4, where the a value for the components is supplied in Table III.

$$N = (p)(a) \quad (4)$$

	Normal	Axial	Pitch
a	1.6	1.2	2.2

Table III. Factors used to estimate data volume.

Recall, the number of parameters in a complete second order model is 28, and in a complete third order model there are 84 (including the intercepts). These values were used to estimate the required data volume, shown in Table IV.

	Normal	Axial	Pitch
2 nd order model	46	32	60
3 rd order model	138	97	181

Table IV. Required data volume based on the number of model parameters.

Therefore, to estimate a second order model for the normal force component a design with at least 46 data

points is required to achieve the specified precision in the presence of the estimated repeatability of the measurement environment. For a complete third order model, a minimum of 138 data points is required. These estimates are conservative because in actual practice the number of parameters in the model that are statistically significant is considerably less. The concepts involved with determining statistically significant terms and model reduction techniques are discussed later in this paper.

It is important to note that the volume of data required is independent of the type of terms in the mathematical model, it is only dependent on the number of terms in the model. This means that with 46 data points, in the example above, a total of 28 terms can be estimated. These 28 terms may include a combination of linear, second order, third order, and higher order. The setting of the independent variables within the design determines the terms that can be estimated.

The estimation of required data volume for a calibration experiment is usually not performed. As a result, it has been generally accepted that the more data that can be obtained, limited only by available resources, the better the calibration result. Resources at research laboratories are never unlimited, and therefore the estimation of the required data volume provides a means to leverage the available resources. The typical OFAT calibration design that has been used at LaRC consists of 729 data points for the estimation of a second order mathematical model. It is obvious that this number of points is an order of magnitude greater than what is required.

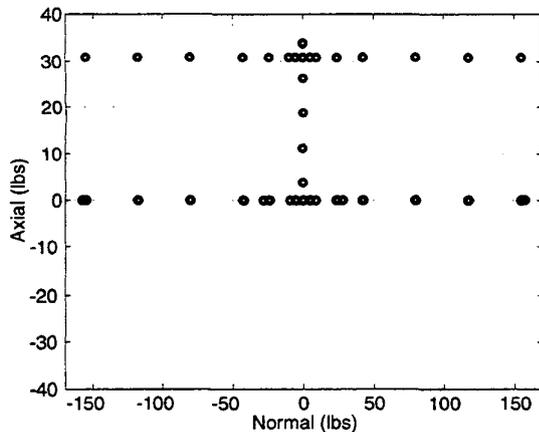
An MDOE approach deviates from the current trend of collecting massive data volume in an OFAT method, by specifying ample data to meet requirements quantified in the design without prescribing volumes of data far in excess of this minimum. The goal is to efficiently achieve the primary objective of the calibration experiment; namely the determination of an accurate mathematical model that meets the specified objectives.

In the application of the force balances to wind tunnel testing, the prediction quality (precision requirement) and prediction risk (inference error risk) are typically not provided by the aerodynamic researcher. Their impact on the ability to answer the research questions under investigation in the wind tunnel test should drive these objectives, and thereby the data volume. This link between calibration required accuracy and the ability to adequately answer the aerodynamic research questions is vital to apply appropriate resources to the balance calibration experiment.

Design Evaluation

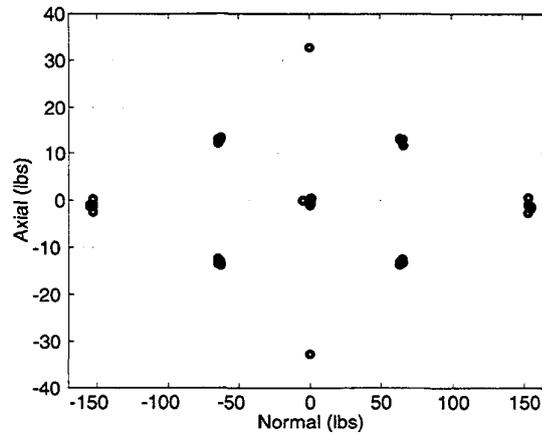
What makes one calibration design better than another? How do we objectively evaluate the prediction capability of a calibration design? Prior to the execution of the design, the quality can be evaluated by a number of techniques. Two of these techniques will now be presented. First, the design space, the region over which the calibration is performed, should be compared to the region of operability. A force balance has a six dimensional inference space, which can be thought of as a hyper-cube and contains all possible combinations of the independent variables. The region of operability is the region in which the balance will be used in the wind tunnel. Since the model is an approximation of the true functional relationship of the balance response to applied load, the design space should be chosen to be as close as possible to the region of operability, because we want to perform that approximation over as small of a range as possible. This implies that a calibration should be tailored to the requirements of a specific wind tunnel entry. Currently, due to productivity constraints, this is not performed and the balance is calibrated over its entire load range. Also, we want the region of operability to lie within the perimeter of the design space to ensure that we do not use our model to extrapolate, which can produce significantly higher errors in the prediction.

A graphical comparison of the inference space is useful in comparing different calibration designs. As an example, using the UT-39A balance, the setting of the independent variables for the normal force and axial force components are provided in Figure 2.



(a) OFAT 729-point design

Figure 2. Normal force and axial force combinations.



(b) MDOE 64-point design

Figure 2. concluded.

A comparison of these figures clearly reveals differences in the design space. In particular, the OFAT design does not include any settings of negative axial force. It has been generally considered that in wind tunnel testing only positive axial force is seen by the force balance, unless a powered model is used for propulsion type research. In fact, when testing the model at positive pitch angles, a vector component of positive lift acts in the negative axial force direction. Extrapolation is required to predict responses in the negative axial force region of the design space, which is clearly undesirable.

The symmetry of the MDOE design can be seen in Figure 2(b). Recall, that the inference space of the force balance is six dimensional, and the figure is a two dimensional plot. The combinations of the six independent variables are also symmetrical in six dimensional space.

The distribution of the standard error of the predicted values is a second method. The purpose of performing a balance calibration and developing a model is to be able to predict a future response of a specific setting of the independent variables. The prediction standard error is a function of the model, the design, and the location of a point of interest within the design space. The reader is referred to Reference 8 for the mathematical derivation of this distribution. It is the goal of this section to provide a qualitative understanding of this concept, which provides insight into the prediction quality of the design prior to execution. It is important to emphasize that this phase does not require experimental data; rather it is an evaluation of the design itself.

There are many advantages in evaluating the calibration design separate from experimental data. Typically,

calibration designs are evaluated by performing successive calibrations of a particular test balance with various experimental designs. While this is quite useful, the performance of the particular test balance and the quality of the calibration system influence the evaluation, which is unfortunate. In other words, the results of the same series of tests could produce different results using another test balance and calibration system. Our goal in evaluating the design is to determine if the prediction capability is adequate over the region of operability.

The value of the standard error of prediction, at a point in the design space, is computed according to the following equation.⁸

$$\text{Var}[\hat{y}(x)] = x^{(m)'}(X'X)^{-1}x^{(m)}\sigma^2 \quad (5)$$

where, $\text{Var}[\hat{y}(x)]$ is the variance in the predicted value, $x^{(m)}$ is a vector which defines the location of the point of interest, m reflects the form of the model, $(X'X)$ is the first moment of the design matrix, and σ^2 is the unexplained variance. Prior to the execution of the design the value of σ^2 is set equal to one, therefore we can perform an evaluation of the distribution of unit standard error. The point of interest is defined by the vector $x^{(m)}$. For a second order model, with six independent variables, $x^{(m)}$ would contain 28 terms computed at the point of interest in the following form,

$$x^{(m)} = [1, x_1, \dots, x_6, x_1x_1, \dots, x_6x_6, x_1x_2, \dots, x_5x_6] \quad (6)$$

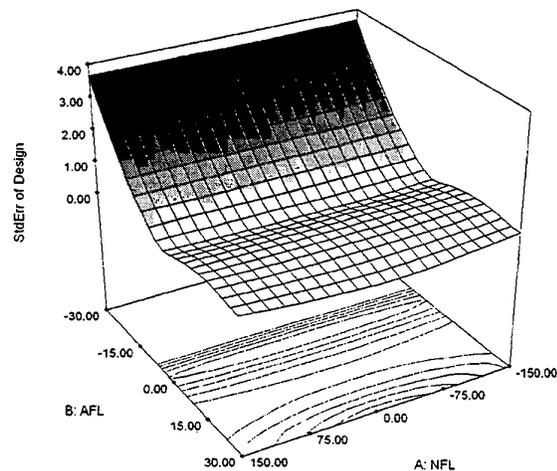
where x_i would correspond to normal force, x_ix_i would correspond to normal force squared, and x_ix_j would correspond to normal force times axial force.

For a given location in the inference space, and a given measurement environment, the quality of the response prediction depends completely on the design matrix, which is a function of the settings of the independent variables and the form of the model.

It is desirable that the calibration design possesses a reasonably stable distribution of the prediction variance throughout the design space. Since, the actual location of prediction within the inference space is not well defined prior to the wind tunnel test; a stable distribution provides insurance that the quality of the prediction is nearly the same throughout the region of interest.

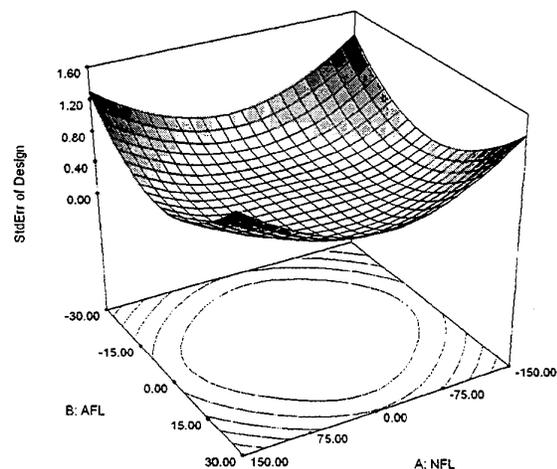
As an example, this evaluation technique is applied to the UT-39A balance. Two experimental designs will be evaluated and compared, the OFAT 729-point and an MDOE 64-point. Shown in Figure 3 are plots of the distribution of unit standard error for the variables of normal and axial. This is a graph of the square root of prediction variance (given in Equation 5) letting the $x^{(m)}$

range over the entire design space. It represents the model's error distribution in multiples of the standard error (square root of the unexplained variance) in the prediction.



(a) OFAT 729-point design

Figure 3. Distribution of unit standard error.



(b) MDOE 64-point design

Figure 3. concluded.

In Figure 3(a), the results of the extrapolation into the negative axial force region of the design space are apparent. As expected, this extrapolation would provide unacceptably high prediction uncertainty.

In Figure 3(b), the stability of the distribution of the unit standard error can be seen by the flat region throughout most of the design space. The near radial symmetry in the contours of constant error depicted in the lower plane are evidence of a desirable property known as rotatability. A design that is rotatable provides the same value of the prediction variance for

all points that are equal distance from the center of the design. The slight error increase in the corners of this design are directly related to the setting of the independent variables depicted in Figure 2(b). Once again this increase in uncertainty is due to extrapolation. In the MDOE design the corners of the design space, which represent all six independent variables set at their maximum values simultaneously, have been considered as unlikely combinations to occur in actual wind tunnel testing, and therefore are not set in the experimental design.

Design Execution

How do we measure or defend against systematic errors that are present in all calibration systems? The three fundamental quality-assurance principles employed during the execution of a formal experiment design are randomization, blocking, and replication. Randomization of point ordering ensures that a given setting of the independent variables is just as likely to be applied early in the calibration as late. If sample means are stable, the point ordering does not matter. However, if some systematic variation (e.g. instrumentation drift, temperature effects, operator fatigue, etc.) causes earlier measurements to be biased low and later measurements to be biased high then randomization converts such unseen systematic errors to an additional component of simple random error. Random error is easy to detect and also easy to correct by replication. Randomization of point ordering also increases the statistical independence of each data point in the design.

Statistical independence is often assumed to exist in the current methods of balance calibration, but systematic variation can cause measurement errors to be correlated, and therefore not independent of each other, as required for standard precision interval computations to be valid. Even relatively mild correlation can corrupt variance estimates substantially, introducing significant errors into estimates of "95% confidence intervals" and other such quality metrics.

Blocking entails organizing the design into relatively short blocks of time within which the randomization of point ordering ensures stable sample means and statistical independence of measurements. While randomization defends against systematic within-block variation, substantial between-block systematic variation is also possible. For example, calibrations spanning days or weeks might involve different operators, who each may use slightly different techniques, or possess somewhat different skill levels. By blocking the design, it is possible in the analysis to

remove these between-block components of what would otherwise be unexplained variance.

Averaging replicates causes random errors to cancel. This includes otherwise undetectable systematic variation that is converted to random error by randomizing the execution order of the experimental design. Replication also facilitates unbiased estimates of what is called "pure-error" - the error component due to ordinary chance variations in the data. These pure-error estimates are critical to evaluating the quality of the calibration model by permitting the fit of the model to the data to be assessed objectively.

This section has presented powerful tools that can be used to design and evaluate an experimental design. These tools are especially useful in determining the resources required to execute the design by estimating the required data volume that is dictated by the objectives of the calibration. Historically, little focus has been applied to the construction of the experimental design, even though its impact on the model has been suspected. With these techniques, designs can be evaluated before experimental data is obtained. Also, objective comparisons can be made between various designs. The principles of how to execute the experimental design have been presented as a tactical measure to defend against systematic errors in the calibration process. The way in which the experimental design is constructed and executed enables the application of sophisticated analyses of the data. These analyses are discussed in the next section.

Response Surface Methods

The application of response surface methodology to the analyses of the experimental data is quite different from current balance data processing. It involves statistical tools used in tandem with the experience of the balance engineer to objectively determine the model coefficients. The number of terms in the model is minimized by eliminating those with coefficients that are too small to resolve with a sufficiently high level of confidence. Also, the total unexplained variance is partitioned into the pure-error and lack-of-fit components. This analysis of unexplained variance provides a method to make objective judgements about the adequacy of the model and the potential for improving the model with higher order terms.

Model Selection

During the modeling process, the model with the fewest parameters is desired. A "good" model is the smallest one that has insignificant lack-of-fit and meets the precision requirement. An analysis of variance (ANOVA) is performed to achieve these objectives. One aspect of the ANOVA is a method to determine the

statistical significance of each model coefficient. The axial force component of LaRC balance 2008 is used as an example of how the model terms are selected. The same procedure applies to the other five response variables.

Table V contains a portion of the ANOVA of the reduced second order model. In the first column is the source of variance, this includes the explained variance (model terms and block effects) and the unexplained variance (lack-of-fit and pure-error). The second column contains the number of degrees of freedom (DF) used to estimate the variance from each source. In the third column the variance, or mean square error (MSE), of each model coefficient is provided in units of microvolts per volt quotient squared. The fourth column contains the F-value, which is equal to the ratio of the variance of each coefficient divided by the residual variance. The right-most column contains the probability that an F-value this large could have occurred due to chance variations in the data (experimental noise). The smaller this probability, the more confidence that we have that the model coefficient is non-zero.

Source	DF	Mean Square	F-value	Prob > F
Block	1	3.384E+02		
Model	6	3.665E+05	160,503	< 0.0001
N	1	1.630E+03	714	< 0.0001
A	1	2.158E+06	945,011	< 0.0001
P	1	4.692E+03	2,055	< 0.0001
R	1	1.209E+01	5.3	0.0252
Y	1	9.539E+01	41.8	< 0.0001
NP	1	1.052E+01	4.6	0.0362
Residual	56	2.284E+00		
Lack of Fit	47	2.697E+00	21.93	< 0.0001
Pure Error	9	1.230E-01		
Total	63			

Table V. Second order reduced model.

For example, probability values of less than 0.05 suggest less than a 5% probability of a chance occurrence due to noise resulted in this regression coefficient. If this probability is less than a threshold value then the coefficient is considered significant and is retained in the model. Note the large F-value and associated low probability of the A term on the axial force response. This is expected since it represents the sensitivity constant of that particular component and there is a strong correlation between the application of axial force and the associated axial force response.

The first step in the model selection process was to perform an ANOVA of a complete second order model. Then, a threshold probability of 0.05 was used to determine the reduced model. All coefficients that

were above this probability were removed. After a model term is removed, the regression is performed again with the reduced model. This enables the best possible fit of the coefficients for the terms remaining in the model to be calculated. This procedure is performed in an iterative manner until all terms in the model are considered significant. The complete second order model was reduced from 27 terms to 6 terms (excluding the intercept).

It is the goal of this phase to minimize the number of coefficients in the model, which in turn lowers the average variance, because each coefficient carries some uncertainty. Once the model has been selected, then an analysis of the unexplained variance can be performed.

One diagnostic technique used to determine the quality of the model fit is to perform a normal probability plot of the residual errors. Figure 4 shows this plot for the second order model. The probability scale on the vertical axis assumes a normal distribution of statistically independent data points. The studentized residuals, on the horizontal axis, are multiples of the residual standard deviation. It is the number of standard deviations that separate the actual and predicted response values at a particular point.

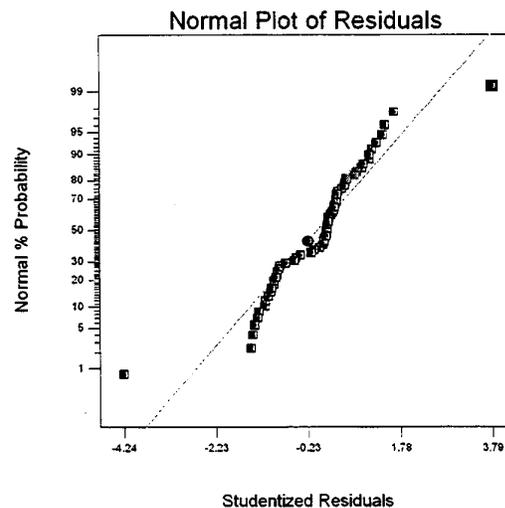


Figure 4. Residuals from second order model.

A pattern in the residuals can indicate a relationship between the balance signals and the independent variables that is not included in the model. In this case, based on structural knowledge of the balance design, it was considered that higher order effects involving normal force and pitching moment were likely. Higher order terms are usually attributed to balance deflection under applied load. In classical solid mechanics equations, these deflection equations are approximated

as cubic, and therefore it is natural that the signals would exhibit cubic behavior. Therefore, five third order terms were added to the model. An additional three second order terms, which were insignificant relative to the residual of the second order model, were also included. The results are shown in Table VI and the corresponding normal probability plot of the residuals is provided in Figure 5. The normal probability plot of the reduced third order model has a better appearance of normality. To further analyze the model, an analysis of the unexplained variance was then performed.

Source	DF	Mean Square	F-value	Prob > F
Block	1	3.384E+02		
Model	14	1.571E+05	558,569	< 0.0001
N	1	5.814E+02	2,067	< 0.0001
A	1	2.103E+06	7,475,452	< 0.0001
P	1	1.154E+03	4,103	< 0.0001
R	1	1.066E+01	37.9	< 0.0001
Y	1	9.711E+01	345.3	< 0.0001
P2	1	2.122E+00	7.5	0.0084
Y2	1	5.074E+00	18.0	< 0.0001
NP	1	8.087E+00	28.8	< 0.0001
PS	1	5.207E+00	18.5	< 0.0001
P3	1	2.559E+01	91.0	< 0.0001
NP2	1	6.439E+01	228.9	< 0.0001
NAP	1	7.439E+00	26.4	< 0.0001
PRS	1	5.914E+00	21.0	< 0.0001
PYS	1	1.959E+00	7.0	0.0112
Residual	48	2.813E-01		
Lack of Fit	39	3.178E-01	2.58	0.0659
Pure Error	9	1.230E-01		
Total	63			

Table VI. Third order reduced model.

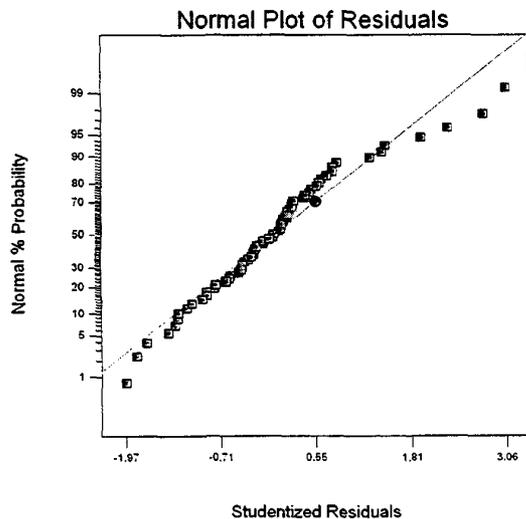


Figure 5. Residuals from third order model.

Analysis of Unexplained Variance

An analysis of the unexplained variance is performed in order to partition the residual error. Typically, in the field of balance calibration, the accuracy is based on the standard deviation of the residual errors obtained by computing the difference between the actual values and the model predicted values. This is expressed as two times the standard deviation providing a 95% confidence interval. The assumption of statistical independence of the data points from an OFAT calibration is not valid. Even relatively mild correlation can corrupt variance estimates substantially, introducing significant errors into estimates of 95% confidence intervals.

More importantly, the total residual error includes two distinct components, lack-of-fit and pure-error. The lack-of-fit relates to the ability of the math model to capture the response of the balance electrical signals as a function of the independent variables. The pure-error is a function of the repeatability of the measurement environment. This includes factors such as the mechanical calibration system, the data acquisition system, the balance instrumentation, the quality of the mechanical interfaces, and the thermal stability of the calibration laboratory. RSM provides a technique to separate the lack-of-fit and pure-error components of the unexplained variance.

First, the pure-error component is computed from the genuine replicates that are performed throughout the calibration experiment. In the case of the MDOE 64-point design, there were a total of eleven replicates, in two blocks. Subtracting the mean value of each block from these replicates provides nine degrees of freedom (DF) to estimate the pure-error. The sum of the squared (SS) deviations from the mean of each block is computed. The mean square error (MSE), variance, can be computed based on Equation 7.

$$MSE_{\text{pure error}} = \frac{SS_{\text{pure error}}}{DF_{\text{pure error}}} \tag{7}$$

Once the pure-error is known, its contribution to the total residual can be determined. This computation involves subtracting the SS of the pure-error from the SS of the total residual as shown in Equation 8.

$$SS_{\text{lack of fit}} = SS_{\text{total residual}} - SS_{\text{pure error}} \tag{8}$$

The MSE of all three quantities (total residual, lack-of-fit, and pure-error) can then be computed using the associated degrees of freedom (DF) and the SS according to Equation 7. The ratio of the MSE of the lack-of-fit divided by the MSE of the pure-error forms the F-value as shown below.

$$F_{value} = \frac{MSE_{lack\ of\ fit}}{MSE_{pure\ error}} \quad (9)$$

This F-value is compared against a critical value of the F-distribution that depends on the degrees of freedom for both lack-of-fit and pure-error, and the specified significance of the test, 0.05 in our case. This 0.05 significance level means that if our measured F-statistic exceeds the critical F-value, we can reject the null hypothesis with 95% confidence. The null hypothesis in this case is as follows: *H₀*: The variance of the lack-of-fit is not significant relative to the variance of the pure-error. If the F-value is greater than the critical F-value then the null hypothesis can be rejected. In this case, it can be stated that we have 95% confidence that the lack-of-fit is significant. On the other hand, if the F-value is smaller than the critical F-value, then we would not reject the null hypothesis, concluding that we are unable to detect significant lack-of-fit with our required 95% level of confidence. This F-test procedure provides an objective method for determining whether or not the model has significant lack-of-fit. Significant lack-of-fit means that the calibration response function does not adequately represent the data upon which it is based.

A summary of the results of the above analysis performed on balance 2008 is provided in Table VII. The table contains the analysis of unexplained variance for both the second order and third order models presented in the last section.

Quantity	Second Order	Third Order
Maximum Response	559.8	559.7
Lack-of-Fit (DF)	47	39
Lack-of-Fit (MSE)	2.6975	0.3178
Lack-of-Fit (sigma)	1.64	0.56
Pure-Error (DF)	9	9
Pure-Error (MSE)	0.1230	0.1230
Pure-Error (sigma)	0.35	0.35
Residual (DF)	56	48
Residual (MSE)	2.2837	0.2813
Residual (sigma)	1.51	0.53
Measured F-value	21.93	2.58
Critical F-value (@ $\alpha = 0.05$)	2.81	2.83
Significant Lack-of-Fit? (@ $\alpha = 0.05$)	Yes	No
units of quantities: maximum response is (microvolts/volt), MSE is (microvolts/volt) ² , sigma is (microvolts/volt)		

Table VII. Analyses of unexplained variance.

The data in this table provides insight into the mathematical model and the physical calibration

system. The one-sigma estimates in the table are computed by taking the square root of the mean squared error.

The lack-of-fit is considered significant in the second order model, and therefore a higher-order model could be used to provide a better fit. After including the third order model terms, the residual was reduced by 65%, and the F-test now supports rejecting the null hypothesis. In other words, the model does not exhibit significant lack-of-fit at a level of 95% confidence.

It is important to realize that the lack-of-fit test is relative to the pure-error. In the limit, as the pure-error goes to zero, the lack-of-fit F-value goes to infinity. The decision to use a higher-order model is also linked to obtaining the required precision.

Areas of Future Research

Future research efforts include the investigation of a D-optimal design approach combined with multivariate orthogonal functions, the implementation of higher-order models, and the expansion of the calibration model to include temperature effects.

Employing a D-optimal approach to the construction of the experimental design has the advantage of selecting points from candidate combinations of the independent variables that are tailored to the physical constraints imposed by the mechanical calibration system. A D-optimal design minimizes the volume of the joint confidence region on the vector of regression coefficients.⁹ To achieve the orthogonality of the regressors, multivariate orthogonal functions can be utilized.¹⁰ Combining these two techniques will enable the construction of new experimental design that will improve the efficiency of the execution of the design.

RSM techniques provide systematic methods for research into better mathematical models. It is common practice to include partitioned coefficients in a balance math model to improve second order model deficiencies. These partitioned coefficients, often referred to as split terms, are more likely higher order terms. Designs will be executed to investigate complete cubic math models.

It is generally known that balance calibration response is a function of temperature. At the present time, all LaRC balance calibrations are performed at room temperature. Few, if any, operate at room temperature in the wind tunnel environment. In some cases, an abbreviated OFAT sequence of loads is performed at elevated or cryogenic temperature. The temperature calibration results are difficult to interpret due to the inability to separate the repeatability of the measurement environment from the actual thermal

effects. A calibration design that incorporates balance temperature as an independent variable has been proposed.

Concluding Remarks

Certain weaknesses in current balance calibration experimental methodology and mechanical systems have been presented. The Single-Vector System enables an MDOE approach to balance calibration that can address these weaknesses. Practical application of these techniques has been emphasized with actual experimental data. The following specific findings are noted:

- 1) The data volume of the experiment is functionally related to the requirements of the balance performance and the precision of the calibration system. Current OFAT designs have an order of magnitude higher data volume than required. Calibration resources can be better utilized if ample data volume is specified, not excessive data volume.
- 2) An experimental design can be evaluated to determine its adequacy to meet the objectives. This evaluation, which does not require experimental data, enables objective comparisons of various designs.
- 3) Randomization of point ordering, replication of design points, and blocking can be used as tactical measures to defend against systematic errors, present in all calibration systems.
- 4) A "good" mathematical model minimizes the number of terms and eliminates those that can not be distinguished from experimental noise with a high level of confidence. Genuine replicates enable the partitioning of the unexplained variance into lack-of-fit and pure-error components.
- 5) Partitioning of the unexplained variance provides an objective method to determine when higher order models are justified. Normal probability plots provide a useful graphical method that aid in determining model adequacy.

These sophisticated and elegant techniques have been routinely used in many fields outside of experimental aeronautics. Application of these methods to balance calibration provides new insight to the calibration process, and an increased quality of force and moment measurements during wind tunnel testing.

Acknowledgements

The authors would like to thank the Langley Wind Tunnel Reinvestment Program who provided funding for this effort, Greg Jones and Wesley Vellines of Modern Machine and Tool Company for their assistance in the execution of the calibration designs,

and Gary Erickson of NASA LaRC for his support in making the UT-39A and 2008 available for this research effort.

References

- 1) Box, G.E.P.; Draper, N.: *Empirical Model-Building and Response Surfaces*. John Wiley & Sons. 1987.
- 2) Ferris, A. T.: *Strain Gauge Balance Calibration and Data Reduction at NASA Langley Research Center*. Paper DR-2, First International Symposium on Strain Gauge Balances. October, 1996.
- 3) Parker, P.A.; and Rhew R.D.: *A Study of Automatic Balance Calibration System Capabilities*. Second International Symposium on Strain Gauge Balances. May 1999.
- 4) Parker, P.A.; Morton, M.; Draper, N.; Line, W.: *A Single-Vector Force Calibration Method Featuring the Modern Design of Experiments*. AIAA 2001-0170, 39th Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 2001.
- 5) Hansen, R.M.: *Evaluation and Calibration of Wire-Strain-Gage Wind-Tunnel Balances Under Load*. NACA Langley Aeronautical Laboratory, 1956.
- 6) Diamond, W.J.: *Practical Experiment Designs for Engineers and Scientists*. John Wiley & Sons. 2001.
- 7) DeLoach, R.: *Tailoring Wind Tunnel Data Volume Requirements through the Formal Design of Experiments*. AIAA 98-2884, 20th AIAA Advanced Measurement and Ground Testing Technology Conference, Albuquerque, NM, June 1998.
- 8) Myers, R.H.; and Montgomery, D.C.: *Response Surface Methodology*. John Wiley & Sons. 1995.
- 9) Montgomery, D.C.: *Design and Analysis of Experiments, 4th edition*. John Wiley & Sons. 1997.
- 10) Morelli, E.A.; and DeLoach, R.: *Response Surface Modeling Using Multivariate Orthogonal Functions*. AIAA 2001-0168, 39th Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 2001.